Rheedea Vol. 6(2) :11–45.1996.

Phylogenetic Research: A Short Introduction

Peter C. van Welzen
Rijksherbarium/Hortus Botanicus, University of Leiden, P.O. Box 9514,
2300 RA Leiden, The Netherlands

Abstract

The purpose of this article is to provide an easy to understand explanation of phylogenetic (cladistic) research, which might compel every reader to give this type of research a try. For this reason the text does not only explain the methodology, it also advices, shows examples and offers a brief manual to one of the computer programs. Topics which are briefly treated are the use of taxa, characters, outgroups, parsimony, optimisation algorithms, computer programs, interpretation of results, phylogenetic classification, and future developments.

INTRODUCTION

It is not an easy task to present a comprehensive summary of the methods of phylogenetic (or cladistic) analysis, which I am invited to write. The best way to start is to erase a few misunderstandings about phylogenetic research.

It is a method to group taxa, different from the classical and the phenetic way of doing this. In fact, it is not even necessary to produce a classification. The result of a phylogenetic analysis is in fact a cladogram or phylogenetic tree, a branching diagram showing the evolutionary relationships between the taxa analysed (e.g. Fig. 1). This cladogram may be translated into a classification.

Phylogenetic research has nothing to do with the species concept of researchers. The species have to be delimited already before the analysis starts. It may happen that after the analysis one is inclined to change the concept of certain species. An interesting philosophical resolution of the problems concerning the species concept is presented by Kornet (1993a, b), which has very nice practical applications.

The phylogenetic school of thought, just like the phenetic school, is a reaction to the classical way of grouping species. In the latter 'school' researchers tend not to use rules, do not discuss the importance of their characters, nor do most of them elaborate on their decisions. Therefore, the results of their research could often not be repeated. Later on I will indicate the difference between the phenetic and phylogenetic type of analysis.

Peter C. van Welzen

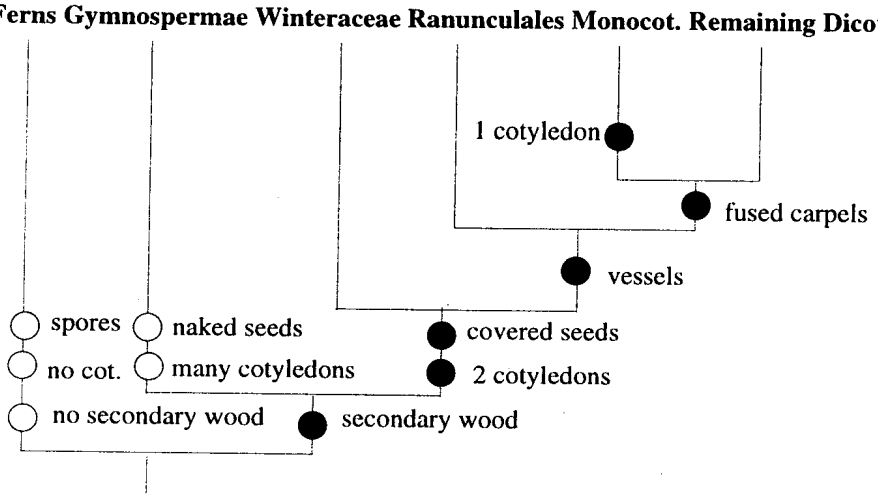


Fig. 1. Symplified cladogram of the higher plants. The black dots represent apomorphies. The opposite plesiomorphies are only shown for three characters (white dots).

The aim of the phylogenetic methodology, initiated by Hennig, a German entomologist (1950, English translation in 1966), is to group species together on the basis of common ancestry. Species which evolved from the same ancestor are grouped together. Characters should display the sequence in which the species evolved. The idea is that the acquirement of new characters in a certain species is retained and passed on to descendent species. The new character then becomes typical for a whole group of species. An example is the presence of an opposable tumb, once originated in an ancestor, now typical for all monkeys and half monkeys. A later ancestor within the monkeys lost its tail and this was passed on to its descendents, the group of monkeys we know as man-apes (humans, gibbons, chimpanzees, gorillas, orang utans). This means that all monkeys evolved from one ancestor, the one with the famous tumb, and within the monkeys the loss of a tail indicates that all man-apes had an ancestor in common. During a phylogenetic analysis we try to find these internested characters in such a way that a dichotomous scheme of species splits emerges, this will be the cladogram.

Fig. 1 shows a cladogram. Let me explain what it shows. Vertically we see a very relative time scale, horizontally there is nothing, just space to seperate the different taxa we are analysing. Time increases from bottom to top, which means that the youngest time is at the top. Furthermore, we see lines and these are connected at their base, the node. From every node two lines emerge. The lines indicate (possible) species (others will argue that only the nodes represent species). At the bottom we see that an ancestor of all higher plants existed, typical for this species was the presence of secondary wood. From this ancestor, following the line to the right, we see a split into the Gymnosperms and the Angiosperms. Typical for the latter are the

2 cotyledons and the covered seeds. The next split results in the family Winteraceae and the ancestor to the rest of the Angiosperms (typically with vessels in its wood). The latter ancestor splits into the Ranunculales and the remaining Angiosperms, etc. The presence of covered seeds, united carpels, vessels, 2 cotyledons, 1 cotyledon are typical for a group of plants. These typical characters are called apomorphies (or synapomorphies when grouping several taxa together and autapomorphies when typical for just one taxon, like one cotyledón for the Monocotyledons). The opposite, here only indicated for the three bottom characters, but present in all, are called (sym)plesiomorphies. The latter do not indicate anything, the absence of vessels is not typical for the Winteraceae, but is also found in Gymnosperms, mosses, algae, etc. and it certainly does not tell us that these groups should be classified together. The same with the tails of the monkeys, the presence of a tail does not indicate that the monkeys had one ancestor, because also other mammals, fishes, birds, etc. have tails. At one time tails may have been important as an apomorphy, perhaps to define the Chordata. Fig. 1 also shows that a character is at one time an apomorphy, after that it becomes a plesiomorphy. Two cotyledons are typical for the Angiosperms, within the Angiosperms it does not indicate any other relationship, but the presence of one cotyledon does, this is indicative of the Monocotelydons. Here we find the main difference with phenetic analyses, the latter do not divide characters into apomorphies and plesiomorphies, but every character is used on every level in the analysis, not just once on one level (not like 2 cotyledons is typical for only the Angiosperms, but it is also as typical for the Ranunculales, the Winteraceae, etc.). The length of the lines in a cladogram does not indicate real time, one part may be a 100,000 years, while another part at the same height in the cladogram may have lasted several millions of years. Also, the inferred evolution in a cladogram is always dichotomous, but in reality polytomies (the splitting into more than 2 species) may have occured. However, a purely dichotomous scheme is the most informative, later on, when real time is added, some branches may appear to be 0 years long and the polytomy is a result. Moreover, it is not necessary that the ancestral species becomes extinct, daughter species may split off from it. The algorithms used in a phylogenetic analysis do not permit splitting off, only splitting up. (Indicative for the continued presence of an ancestral species may be the absence of autapomorphies in one of the descendents.) In literature an elaborate discussion exists how phylogenetic trees can be inferred from a cladogram. We will just use the cladograms as produced by the algorithms.

In most published analyses usually taxa of the same rank are analysed (e.g., only species, genera, families). In fact this is not necessary, because the circumscriptions of these taxa are very variable and what is considered genus level in one family, will only be species level in the other.

In the following chapters I will give a step by step prescription how a simple phylogenetic analysis can be performed. I will try to refrain from dogmatism. Phylogenetic research cannot always be applied. In most cases work on local floras is not suitable for cladistic analyses, nor will it be possible to have satisfactory results if the differences between species are very slight or if much parallel evolution occurred. Readers, who like to have a broader view, are referred to the book of Forey *et al.* (1992) and to Wiley (1981).

Peter C. van Welzen

INPUT

The input in a cladistic analysis is formed by the taxa and the characters. Several restrictions are present as to which characters and which taxa we can use.

Taxa should be monophyletic. This means we should only analyse groups which are derived from one ancestor and with all the known descendent species of this ancestor included in the analysis. We should not analyse paraphyletic groups (groups also derived from one ancestor, but with not all descendents included) or polyphyletic groups (groups derived from more than one ancestor). Some examples: monophyletic groups in Fig. 1 are Rest of Dicotyledons, Monocotelydons, both together!, both together with the Ranunculaceae, etc. Paraphyletic groups will be the Dicotyledonae (should include the Monocotyledonae) and a polyphyletic group will be the Mosses together with the Red Algae, or bats and birds, or succulent Euphorbs and cacti.

Of course the difficulty will be to know before the analysis starts which groups will be monophyletic. Only after the analysis has finished one can be more or less certain of the monophyly of a group. We can only make a calculated guess about the monophyly of a group. Important will be whether or not our group has very typical characters (which are probably apomorphies for the group), or that the group is very homogeneous and has a typical set of characters (this means that every separate character is not typical for the group, but the combination of the characters is typical). If we suspect that other taxa might have been derived from our group, then it will be better to include these in the analysis. If later on it appears that they are not derived from our group, then they can be deleted again.

Characters are difficult too. The best essays on characters are written by Pimentel and Riggins (1987) and Stevens (1980, 1991), although the latter may make you feel desperate. Before we continue we should differentiate between characters and character states. A character is a quality of a taxon, the character state the way in which the quality is shown by the taxon. A few examples will clarify this. A character may be flower colour, the states are the actual colours, like red, blue, green, etc. The size of a leaf is a character, the state is the actual length. The characters may be hierarchic, leaves with or without hairs, if with hairs: which type of hair, simple or stellate.

Characters can either be discrete or continuous. The latter cannot be used in a cladistic analysis, because we cannot distinguish character states. Usually no numerical characters, like lengths, ratios, etc. can be used. We can only use discrete, usually qualitative characters, like flower colour, or number of petals. Although quite a few of these so-called qualitative characters can be almost continuous too, like all shades of pink between plants with white and red petals; Stevens (1991) provides a good overview of these kinds of characters.

Character states should not only be exclusive (e.g., 5, 6, 7 petals are three good states; 5--7, 6--7, and 7 are not), they should also be homologous. This means that petals of the buttercup (*Ranunculus*) cannot be compared to petals of plants of other families, because they are not homologous. Homologous structures are either derived from the same ancestral struc-

ture (Fig. 2b) or they are modifications of each other (Fig. 2a). Examples are the wing of a bird, the arm of a human and the wing of a bat. However, if we only compare birds with bats then the wings are not homologous (try to figure out this one, the answer is in monophyly). Usually it is very difficult to recognize homology, therefore we use the principle of parsimony (which we will refer to later again in a different sense). Parsimony means that we assume as few changes as possible. Thus we assume homology unless there is proof of the opposite (like the petals of the buttercups). But be careful, too many mistakes in homology assumptions will render your analysis useless.

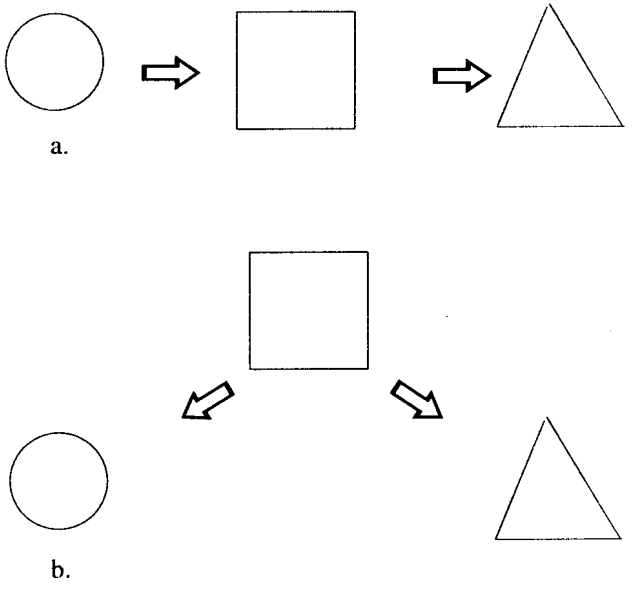


Fig. 2. Character transformations, the different states are homologous. a. They may have been transformed from one another, or b. they are derived from the same anscestral form.

A few extra remarks about characters. Every state in a character receives a number, e.g. 0, 1, 2, etc. This way up to 10 character states per character can be distinguished. If a character only has two states, then it is called a binary character, otherwise it is a multistate character. The numbers may be given at random, then the character is not ordered. If we like to show order, then the states must be ordered in a transformation series whereby the consecutive states get consecutive numbers (e.g. state 3 means that this state developed from either state 2 or state 4). When we have order we still do not have a starting point, if we assign this, for instance to state 2 (see Fig. 3), then we know that the evolution started with state 2, the ordered character is now polarized. Independently state 1 and 3 developed from 2, and finally from 3 state 4 developed.

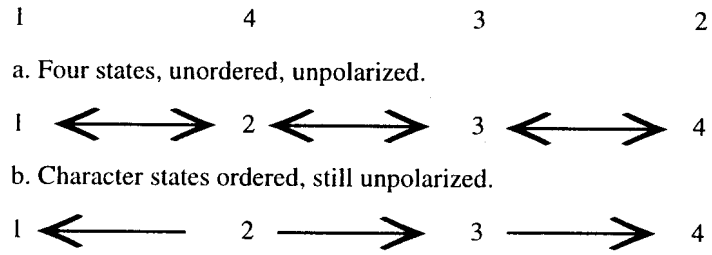


Fig. 3. Ordering and polarisation of characters.

Often it is difficult to divide characters and character states, because different ways are possible. When you are making a list of possible characters and their states, also list alternatives. You can always change to these alternaternatives if the results of the analysis are not optimal.

There are three cases in which we cannot note a character state for a certain species, then we use question marks instead. A question mark indicates that all character states are possible. Usually, after the analysis, we will be able to say what state should have been present in the taxon. This is a hypothesis the cladogram offers and which can be used to falsify the cladogram.

a. Characters are lacking because of the scarcity of the material, for instance the fruits are unknown, then use a ? for this taxon in all fruit characters.

b. A taxon exhibits more than one character state. If many taxa do this, then the character is useless, not discrete, but if only one or two taxa show so-called polytypism, then we enter a ? for the polytypic characters of these taxa (elaborately discussed by Turner & Kornet, ms. and chapter 3.2.1.1 in Turner, 1995).

c. With dependent characters we use a ? for those taxa which lack the principle character. Like with the hairs on the leaf, for those taxa which lack leaves we cannot tell whether they had simple or stellate ones, then we use ? for the hair characters and other leaf dependent characters.

Now a short example of how to construct a data matrix (for a more elaborate treatment on coding see Pimentel & Riggins, 1987). We have 4 taxa and their descriptions. In the matrix we enter the taxa horizontally and the characters vertically. I will not presume to know any order in the characters, nor any starting point (unordered, unpolarized).

Phylogenetic Research: A Short Introduction

Outgroup

Tree, leaves ovate, 12-20 cm long, margin entire. Sepals 5. Petals orange. Fruit a berry.

A

Parasitic herb, scales ovate, 4-8 mm long, margin crenate. Sepals 4. Petals yellow. Fruit a berry.

B

Herb, leaves ovate, 5-10 cm long, margin crenate. Sepals 4. Petals white. Fruit a capsule.

C

Tree, leaves elliptic, 7--15 cm long, margin entire. Sepals 4. Petals blue. Fruit a capsule.

D

Tree, leaves elliptic, 6-13 cm long, margin entire. Sepals 4. Petals green. Fruit a capsule.

N.B: The use of an outgroup will be explained below.

We can discrimate the following characters:

1. Habit.
   - 1 = tree
   - 2 = parasitic herb
   - 3 = herb

or alternatively 2 characters:

1a. Habit	1b. Parasitism
1 = tree	1 = no parasite
2 = herb	2 = parasite

2. Leaf shape
   - 1 = ovate
   - 2 = elliptic

N.B : We do not specially distinguish between leaf and scale, because the latter is part of the parasite syndrome.

(We can use leaf size, because the scales are much smaller than the leaves, but this will give too much weight to the parasite, therefore we will not use this character).

3. Leaf margin (can be continuous when leaves laxly crenate!)
   - 1 = entire
   - 2 = crenate

4. Sepal number (more or less a continuous character)
   - 1 = 5 sepals
   - 2 = 4 sepals

5. Petal colour
   - 1 = orange
   - 2 = yellow
   - 3 = white
   - 4 = blue
   - 5 = green

N.B: This character might not be very useful in our analysis, because every species has a different state, none of the states will group species together, which in fact is our purpose.

6. Fruit type
   - 1 = berry
   - 2 = capsule

Matrix:

	1	2	3	4	5	6
outgroup	1	1	1	1	1	1
A	2	1	2	2	2	1
B	3	1	2	2	3	2
C	1	2	1	2	4	2
D	1	2	1	2	5	2

Alternative matrix:

	1a	1b	2	3	4	5	6
outgroup	1	1	1	1	1	1	1
A	2	2	1	2	2	2	1
B	2	1	1	2	2	3	2
C	1	1	2	1	2	4	2
D	1	1	2	1	2	5	2

Now we have homologized quite a lot, the character herb is regarded in three instances to be the same (alternative matrix) or only in two instances, while the habit herb is homologous with the habit shrub. The same with the shapes. The flower colours are all regarded to be homologous, while often the chemical pathways are very different for these colours. Finally we have said that the fruits are homologous, not only that one berry is equal to another berry, but also that drupes and berries are derived from each other.

CLADOGRAM BUILDING, ALGORITHMS

Usually we need to create a matrix in order to analyse our data by computer. In most cases the data are too complicated to do the analysis manually. But it is often very instructive to show how it can be done that way. Consider the following matrix:

	1	2	3	4
A	1	0	0	0
B	1	1	0	0
C	1	1	1	0
D	1	1	1	1

All taxa have a different combination of character states and each can be recognized. The characters are binary (2 states), the zero is often used to indicate absence.

We can now start to look for internested characters, like with the monkeys. This means that the characters present in all taxa occurred first, and the ones present in only one or two occurred later. Fig. 4 illustrates the different steps. Character 1 is present in all taxa, thus this is considered to be the oldest character (N.B. it can be older than this group! and be present in more taxa). Now we assume that the ancestor of ABCD had character 1. Next we see that character 2 is present in most taxa except A. This means ABCD split into A and BCD, with character 2 only present in BCD. Next, within BCD, we see that character 3 is present in C and D, which means that BCD split into B and CD. Finally C and D split, with D having character 4.

Phylogenetic Research: A Short Introduction

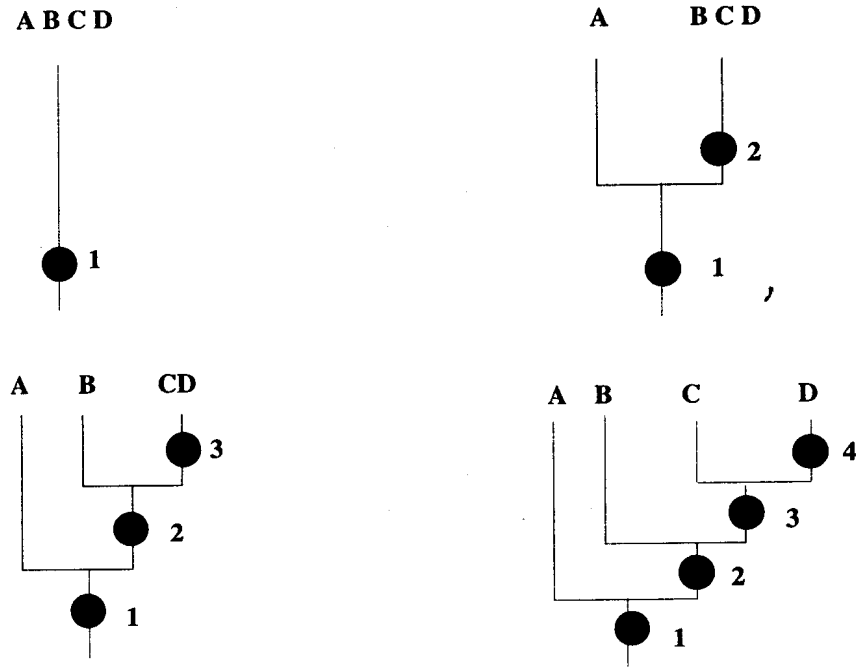


Fig. 4. Stepwise construction of a cladogram based on the matrix provided in the text.

However, although I stated that the characters were not ordered, we did use order, we assumed that 0 is more primitive in comparison to 1. But absence can be a further, less primitive (usually called derived) character state. If we alter the matrix with this in mind, then we obtain the following (Fig. 5):

	1	2	3	4
A	1	0	0	0
B	1	1	0	0
C	1	1	1	0
D	1	1	1	1

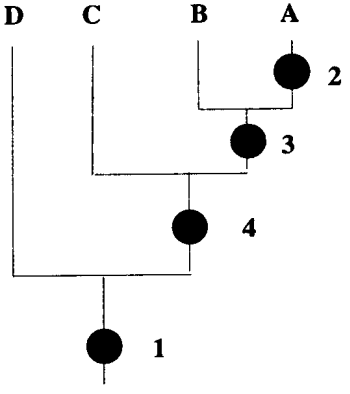


Fig. 5. Alternative cladogram, based on a different interpretation of the data of Fig. 4. Now absence is considered to be apomorphic.

This cladogram is the opposite of the first one. In the two cladograms we see 4 changes. In fact there is no way in which we can choose between the two. What we need is a point of reference, which can tell us which character states are primitive and which are not. Three options exist for this purpose.

a. We can use fossils, but if the fossil record is too imperfect, the results will be very untrustworthy; on the other hand if the fossil record is complete we do not need a phylogenetic analysis.

b. We can use ontogeny. The assumption is that during ontogeny the evolution is briefly repeated (human embryos at one stage do possess a tail). However, especially with plants, we do not possess enough data.

c. The most widely used method is to use the so-called outgroup. We add an extra taxon at the base of the cladogram. Character states which are present in the outgroup and the group we are analysing, referred to as ingroup, are considered to be primitive, and the alternative character states in our ingroup are more derived. Thus, for instance if the outgroup has most character states in common with A then we know that the cladogram in Fig. 4 is correct; otherwise if most states are in common with D, then the cladogram in Fig. 5 is correct.

Choosing an outgroup is often not as easy as it sounds. The outgroup should preferably be the sister group of our ingroup, which means that the two together form a monophyletic group. Sometimes, this choice is apparent due to shared special characters. If not apparent, then we should take care not to use a group which is relatively more derived, otherwise we still select the incorrect cladogram. Then we perhaps better select a more distant group of which we are certain that it has less derived characters (although this may cause the problem that the outgroup has different states, not present in the ingroup, and will not help to polarize the states of the ingroup, it will be like using a fish to polarize plant data). Much can be said about how to choose an outgroup. The best thing to do is to try several taxa as outgroup and do not add only one per analysis, but add at least two or three, this will also help to understand changes at the base of the cladogram (to which I will refer later).

The use of the outgroup rule (those states present in ingroup and outgroup are plesiomorphic = primitive) has to do with parsimony. To assume that shared states are plesiomorphic is far more economic than to assume that they are apomorphic. The latter means that the shared character state developed independently twice, once in the outgroup and once in the ingroup; if we assume plesiomorphy for shared state then the state only changes once to a new apomorphic form somewhere in the ingroup.

Also the final choice of a cladogram has to do with parsimony. If we have three taxa then we can make 3 different fully resolved cladograms (Fig. 6), with 4 taxa we can create 15 different cladograms, with 5 a 125, etc., with 10 taxa over a million different trees or cladograms can be made. Only one of these trees will show the correct evolution. We now use the

methodological tool of parsimony and select the cladogram with the lowest number of character changes as the correct cladogram. Parsimony allows us to make a choice, we do not really assume that evolution always occurred parsimoniously. However, as soon as we make a classification or do something else with the cladogram, we indeed use the idea of a parsimonious evolution. One of the reasons to use the most parsimonious cladogram is that this cladogram is the easiest to falsify. (Philosophical way of reasoning of Popper; however, those used to cladistic analysis will not readily know when their cladogram is falsified, whether this is already the case when a different interpretation of one character occurs or when a completely new topology of the tree is constructed.)

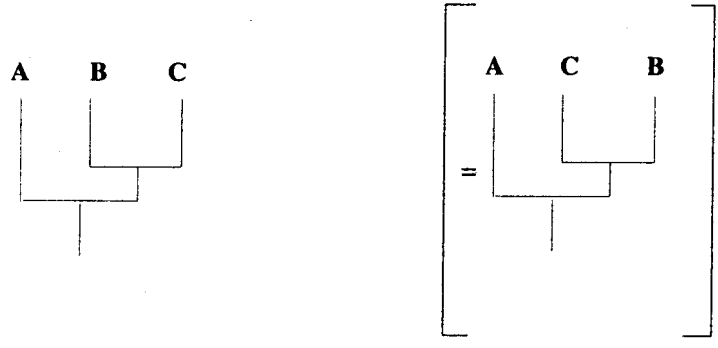


Fig. 6. The three different, fully dichotomous cladograms which can be constructed with three taxa. Alternatives on the same branch (second cladogram) do not count, nor polytomous trees (bottom).

Peter C. van Welzen

Now we know that we need an outgroup, an ingroup (the group we want to analyse) and parsimony to create a cladogram. The cladograms in Fig. 4 & 5 are both as parsimonious, but if we add an outgroup which has character 1(0), 2(1), 3(1), 4(1), and we apply this to both cladograms (Fig. 7) then we see that the second cladogram becomes more parsimonious, therefore we select this one as our most parsimonious cladogram. (The outgroup has most characters in common with D, therefore cladogram 2 is preferred. Exception is the first character, which is different from the taxa A-D. This means that character 1(1) is the synapomorphy for our ingroup.) Please note, that as soon as an outgroup has been assigned, all binary characters are ordered and polarized. Also note that because we place the character changes as parsimonious as possible in the tree, the first tree has an empty branch (the branch BCD is not supported by a character state), which means that in fact we have an unresolved polytomy.

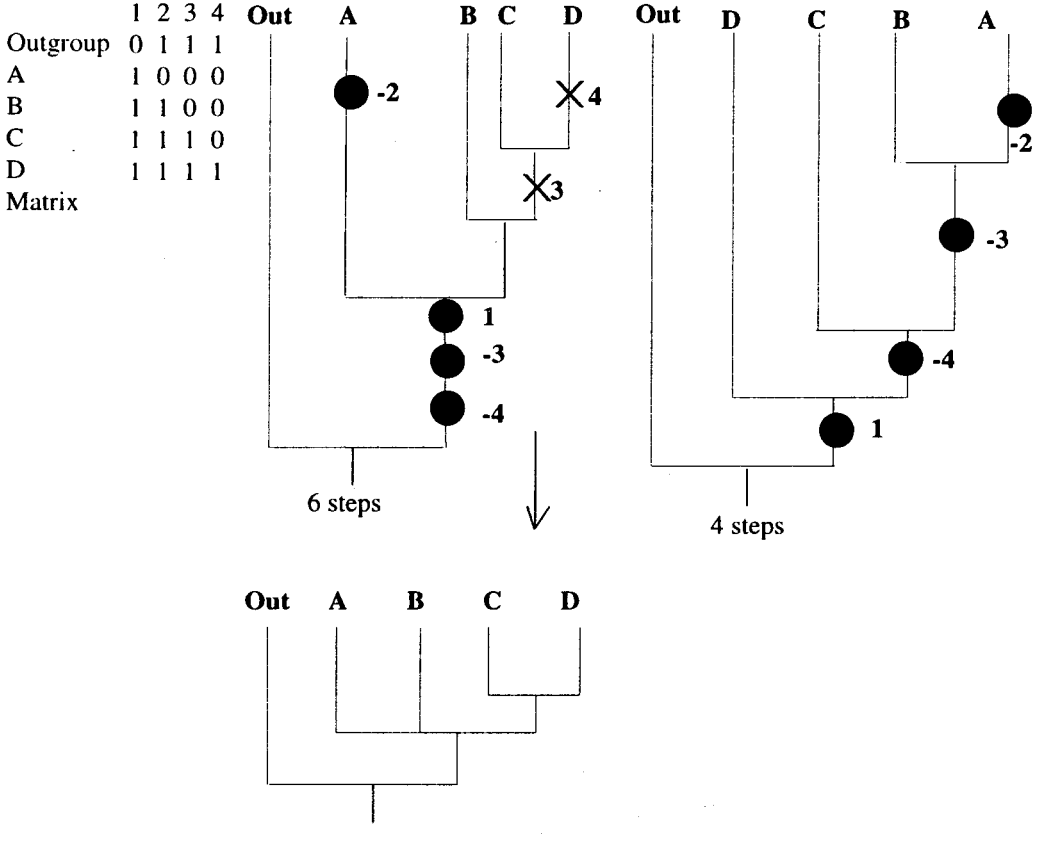


Fig.7. The data of Fig. 4 and 5 revisited. The addition of an outgroup favours the cladogram of Fig. 5 as the most parsimonious cladogram.

Fig. 7 also shows something else. So-called reversals are present in Fig. 7a. A character changes and higher up in the cladogram changes back to the original state. Reversals are one type of homoplasy. The other type is a parallel development (e.g. the thorns of Cactaceae and several Euphorbiaceae). Homoplasy is the result of incorrect assumptions of homology, apparently in Fig. 7a not all presences of a character have originated at the same time and are therefore not homologous. However, we have little or no biological means to distinguish between the two types of presence and we were therefore compelled to consider them to be the same homologous character state. Homoplasy is what makes phylogenetic research often extremely difficult, a high amount of parallel developments and reversals will make it impossible to find a single reliable parsimonious cladogram.

This also explains why it is difficult to estimate how many characters are needed to perform an analysis. In principle we need the (number of taxa - 1) binary characters if these all point to other inclusive groups and do not show (much) homoplasy. If homoplasy is present we need more characters, to compensate for one case of homoplasy we need two non-homoplasious characters. Of course we need less characters if we have multistate characters. Usually, people like to use as many characters as possible. Often this results in using to many ill-defined characters with often arbitrary states. Be careful, I found that it pays to start with a few characters which are trustworthy (show discrete states and no polytypism) and to add more and more characters in later analyses till a single (or usually and unfortunately several) cladogram(s) are found.

One more item has to be introduced before the computer analysis can start. The computer has to know in which way character changes have to be interpreted, this is called optimisation. Of course the optimisation has to be performed in the most parsimonious way, if two sister species have the same character state, then this did not change in each independently (2 changes), but once in their ancestor. Four main types of optimisation are used, although the first 2 are usually not applied:

-reversals not permitted (Camin-Sokal parsimony)

- parallel developments not permitted (Dollo parsimony; a relaxed form of this is often used in the analysis of molecular data) reversals and parallel developments permitted, but:

- characters ordered (Wagner or Farris parsimony)

- characters unorderd (Fitch parsimony)

In the latter two every change is possible, e.g. from character state 4 to 2 or from character state 1 to 2 and vice versa, only the number of changes is counted differently. With Wagner optimisation a change from state 4 to 2 (or vice versa) is regarded as 2 steps, while it (and every other change) is only 1 step in Fitch parsimony. In both types of optimisation a change between consecutive numbers, e.g. 1 and 2 is regarded as 1 step, therefore only multistate characters may provide different cladograms with the two optimisation methods. One more item chould be mentioned. In Fig. 8 it is shown that homoplasy can be resolved in two different ways if a certain character state is present in two neigbouring clades, which were

split off one after the other. Then the character can be optimised as a parallel development (called Deltran - delayed transformation - in the computer program PAUP, Swofford, 1993) or as a reversal (called Acctran - accelerated transformation). In the two solutions the number of steps remains the same, therefore the computer cannot resolve this for us (only artificially by telling that it should use acctran or deltran), we have to use our biological knowledge to interpret this puzzle (e.g. the gain and loss of a structure might be preferred over two independent gains: acctran; or, in case of the reduction of a petal, its parallel disappearance, deltran, might be favoured).

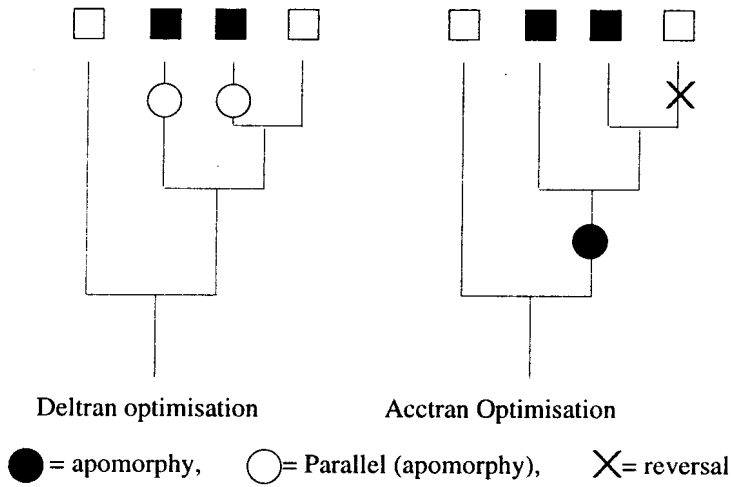


Fig. 8. Different ways of optimisation of a character. Both ways result in the same number of steps (character changes).

Now we are ready to use the computer, we have our taxa, outgroups, character matrix (with characters preferably unordered and unpolarized), and optimisation method. Several programs are available, but this is computer type dependent.

On the Macintosh the use of the programs: Macclade (Maddison & Maddison, 1992; for entering data) - PAUP (calculation of cladogram) - Macclade (interpretation results) is advised.

On DOS machines the programs: text editor (entering data) [- PeeWee (Goloboff, 1993b; mixing datamatrix)] - Hennig86 (Farris, 1988; calculation of cladogram) - Word processor/text editor (interpretation results) are advisable.

Most of these programs are relatively cheap (c. 50-100 US $). The Macintosh programs are most user friendly. Unfortunately, the commonly used program, Hennig86, is very user unfriendly and for this reason an appendix with a short series of commands most used in Hennig86 is provided.

For input one can use PAUP or Hennig86, but one has to know the 'Nexus' format for PAUP and one has to type the complete data matrix in one time correctly for Hennig86.

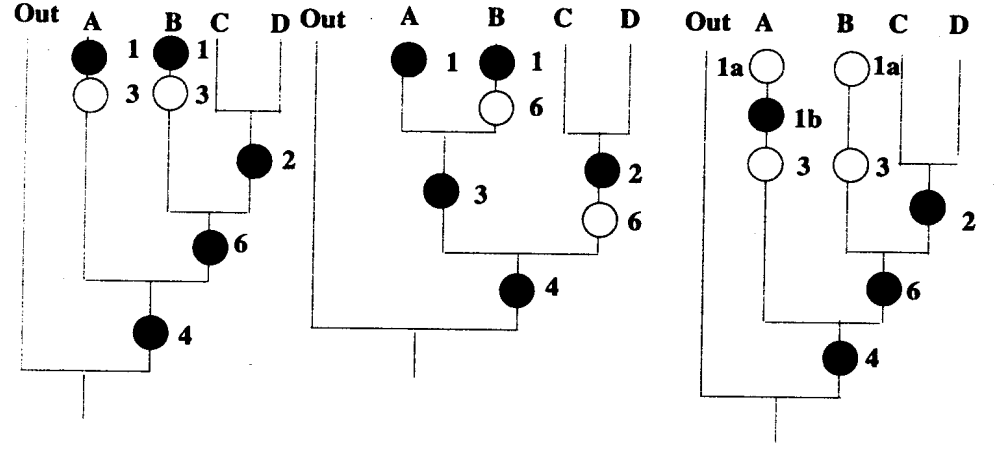
Therefore it is easier to use Macclade to input the data for PAUP (save the matrix in Macclade, start PAUP and use the Macclade matrix as input for PAUP). For Hennig86 it is best to make a batch file with a simple text editor (or save the text as an ASCII-file if a wordprocessor is used). The batch file, which just contains a series of commands, is shown in the appendix and can be used as input in Hennig86 (next to direct commands in the program itself).

If not many taxa are analysed all possible trees can be calculated and one is certain that all most parsimonious trees will be found. For up to 10-12 taxa (dependent on the power of your computer) this is the best possibility (in PAUP command **alltrees**, in Hennig86 more or less command **ie*;**). With more taxa this will cost too much computing time. With up to about 20 taxa a branch and bound algorithm is used, this also ensures that all most parsimonious trees are found (**branch-and-bound** option in PAUP and more or less the combination **mhennig*;** and **bb*;** in Hennig86). With even more taxa a heuristic search has to be performed, of which we are not certain that all or the most parsimonious tree is found. With heuristic searches an initial cladogram is built after which branches are disconnected and rebuilt in different places. The initial cladogram is very important, if the data are complex and the initial tree is very different from the most parsimonious one, then the latter will not be found. The initial tree is dependent again on the sequence of the taxa in the matrix. For this reason it is advis`able to randomise the sequence of taxa and to do this several times, each time producing other initial trees. On all of these branch swapping (different algorithms exist) will be performed and the most parsimonious trees will be saved. In PAUP the **heuristic search** has to be combined with the option **10 (or more) random additions** and the **tree bisecting and reconnecting** algorithm. In Hennig86 this is still the combination of the commands **mhennig*;** and **bb*** ;. To randomise the data for Hennig86 the program PeeWee (Goloboff, 1993b) can be used, this produces several random trees, which can be fed into Hennig86 as input. One more thing, PAUP considers all characters to be unordered (Fitch optimisation), while Hennig86 regards them as ordered (Wagner optimisation). The Hennig86 command **ccode -.;** makes sure that all characters become unordered. Both programs regard the first taxon to be the outgroup, but with separate commands another or more taxa can be defined as outgroup.

Finally, after the programs tell you that the most parsimonious tree has been found (unfortunately, usually many more than one), we can start to interpret the character changes in the cladogram.

INTERPRETATION OF RESULTS

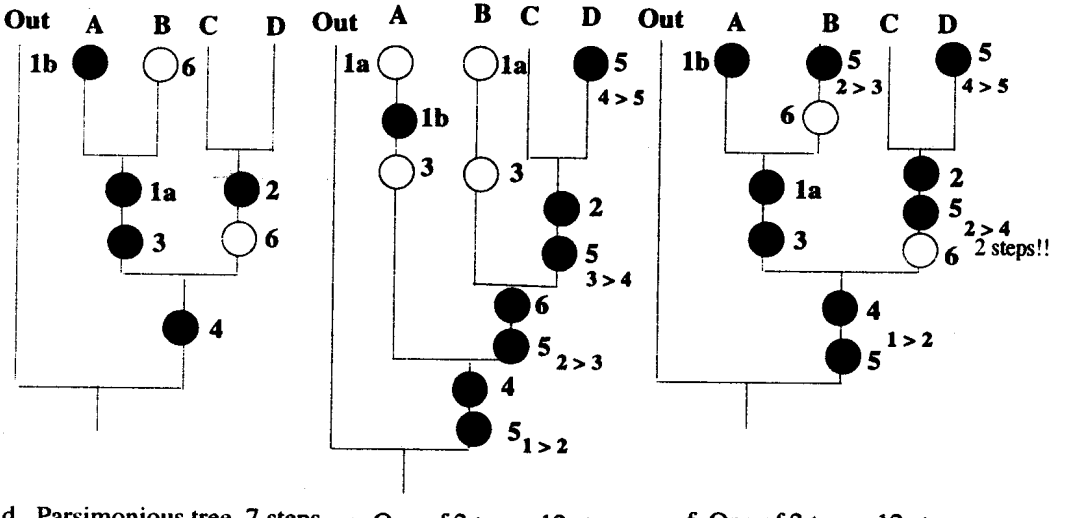
Up to now much theory has been provided and most of it will still remain a black box. Therefore, an appendix has been added, which analyses the results of the two alternative datamatrices we constructed from the descriptions of species A-D and an outgroup. The appendix also serves as a manual for the program Hennig86 (Farris, 1988). If we use all characters as unordered two cladograms are found for the first version of the matrix (Fig. 9a, b), whereby the parasitic habit is regarded as a separate character state of the first character. Both cladograms have 11 character changes and both show one case of homoplasy. Character 3 (leaf margin)



a. One of 2 trees, 7 steps. (character 5 ignored)

b. One of 2 trees, 7 steps. (character 5 ignored)

c. Less Parsimonious,8 steps (character 5 ignored)



d. Parsimonious tree, 7 steps. (character 5 ignored)

e. One of 2 trees, 12 steps. (character 5 ordered).

f. One of 2 trees, 12 steps. (character 5 ordered)

Fig. 9. Different analyses of the datamatrix constructed as example in the text. a and b: Results of first alternative matrix; c and d : Results of second matrix ; e and f : Results of second matrix with character 5 ordered.

shows either a reversal (see fig. in appendix) or a parallel development (Fig. 9a), this is unimportant for the parsimony of the cladogram. In Fig. 9b character 6 (fruit type) shows homoplasy, here too it can be seen as a reversal (see appendix) or as a parallel development (Fig. 9b). In both cladograms character 5 has been omitted, because this can be optimised in many ways, whereby in each instance 4 changes are needed. Characters which show a different state for each species are useless when unordered and are better omitted from an unordered analysis (but see also below). If we analyse the second version of the matrix, with the parasitic habit as a separate character, then only one cladogram is found, Fig. 9d (the same tree as in Fig. 9b). The original alternative (Fig. 9a) is now one step longer (Fig. 9c). This change is brought about by our recoding of the characters, in this case character 1a and 3 support the union of A and B, while only character 6 is contradicting this. In the first two cladograms it was only character 3 supporting the union of A and B, while character 6 united B, C and D, which rendered the analysis inconclusive. In the appendix one special item is shown, character 5, the petal colour, different for every taxon, is turned into an ordered character. Then the same two trees (Fig. 9e, f) are found as in our inital analysis (Fig. 9a, b). In fact this is caused by the linear ordering of character 5 (as shown in Fig. 9e and f). The more symmetric tree (Fig. 9b, d, f) is not favoured anymore, because, although two changes, in character 1a and 3, still favour the union of A and B, it is equalised by a change in character 5 in the group CD where the change from state 2 to 4 counts as 2 steps. In the asymmetric tree (Fig. 9a, c, e) no extra steps are necessary in character 5, but here is one more case of homoplasy (character 1d), which renders both trees as parsimonious again. The problem is caused by the ordering of character 5. It now becomes useful for the analysis, but we only ordered it in a linear way, while the symmetric cladogram indicates that it should have a branched ordering. The latter can be done by not using one character, but by using one character per branch. The branches are from the outgroup to A and B and from the outgroup to C and D. We can re-order character 5 (if we are certain that the symmetric tree has to be favoured) as follows:

	5a	5b	(old 5)
outgroup	1	1	1
A	2	1	2
B	3	1	3
C	1	2	4
D	1	3	5

In the first branch (outgroup, A, B) we use for C and D the state which this branch has in common with the outgroup-A-B branch, which is the state of the outgroup, in the second branch we do the same with A and B (n.b. do not use state 4 and 5 here, then you add extra steps!). If you modify character 5 this way in the matrix and you analyse the results again, you will find that the symmetric tree will be favoured. In this case we have been manipulating the results, because it is often very difficult to suggest any order a priori in the different states. The above way of ordering is introduced here to complete the picture of character coding.

One last remark, the characters at the base of the cladogram, the root, are difficult to interpret. We now interpreted the presence of 4 as an apomorphy for our ingroup. In fact, it is just as parsimonious to see the absence of 4 as an apomorphy for the outgroup (and no apomorphy for the ingroup). A second outgroup, outside the rest, may help here, if it has the same character state as the first outgroup then the interpretation in Fig. 9 for character 4 is correct.

VALUE OF THE RESULTS

It is often difficult to have faith in the results of a cladistic analysis. The choice of an outgroup may not have been correct, the ingroup might not be monophyletic, and the delimitation of characters and character states may have been arbitrary at times. Moreover, a first analysis may have resulted in many parsimonious trees, after which several characters were changed resulting in one or a few trees after new analyses; then one may feel that the results have been highly manipulated. In addition, it may well happen that a tree, one step longer, is much more satisfactory than the most parsimonious one. In other words, how confident can we be in the results of the analysis. For this purpose several indices and statistical methods have been devised (e.g., among the latter jack-kniving and bootstrapping, not explained here, bootstrapping is often used in molecular analyses). The most well-known indices are the consistency index (ci), retention index (ri) and rescaled consistency index (rc).

The consistency index indicates how much homoplasy is present in the cladogram. The index can be calculated per character or for the whole tree. Its maximum is 1 (no homoplasy, perfect fit of characters) and its minimum should be 0 (maximum homoplasy, no fit of the characters, everything developed parallel), but this value cannot be reached. The ci per character or per tree is the minimum number of state changes necessary divided by the number present in the character or in the tree. A few examples, if a character has two states, then the minimum number of changes is 1; if the tree contains only one change in this character then the ci is 1/1=1; if the tree contains 2 changes, then the ci is 1/2=0.5. If a character has 5 states then the minimum number of changes is 4 (total number of states-1), when the tree has 5 changes then the ci=4/5=0.8 (almost no homoplasy), however if there are 12 changes, then the ci=4/12=0.33 (quite a lot of homoplasy). The ci is directly linked to the parsimony of the tree, the trees with the lowest amount of changes (most parsimonious trees) have the highest ci.

The retention index is slightly more difficult. It indicates if a character or tree contains synapomorphies (this together with an indication of homoplasy). If a character only has autapomorphies (autapomorphies are only typical for one terminal taxon, they do not group taxa in the cladogram), then the ri will be 0. If it only contains synapomorphies without homoplasy the ri=1. The ri = (maximum changes - changes in tree)/(maximum changes - minimum changes). The maximum changes can be calculated by regarding every apomorphic state to have been independently developed in all taxa, e.g. if the change is from state 0 to 1 and state 1 is present in 4 taxa, then the maximum number of changes is 4 (4 times a change from 0 to 1). The minimum number of course is 1. Then, if the tree contains 2 changes

(ci=0.5), the ri = (4-2)/(4-1)=2/3=0.67 (which means there are synapomorphies and this character helps to resolve the tree). However, if the number of changes in the tree is 4 (ci=0.25), then the ri=(4-4)/(4-1)=0/3=0 (only autapomorphies, no resolution of the tree).

The rescaled consistency index is ci * ri. This rc, like ri, varies between 0 and 1. It is used by Farris in his computer program Hennig86 for weighting characters (see below).

The analysis may result in not one parsimonious tree, but in several to (really) many trees, all as parsimonious. Then one likes to select a single tree. Several options are available.

1. Consensus trees followed by character evaluation and re-analyses.

2. Character weighting and re-analyses.

3. Different selection criteria.

4. Alternative methods.

ad. 1. From the trees found to be most parsimonious several types of consensus trees can be calculated (strict, semi-strict, Majority, Nelson, etc.). The conflict in the trees will be shown as polytomies (or, in case of majority consensus, as percentages present among the trees). The consensus tree does not represent a true cladogram and cannot be used for purposes like geographic analyses or classifications. It is only a summary of conflict between the different trees. Next we can see which characters are causing the conflict and these characters we can change (delimit different states, combine or separate characters, or add new characters or just ignore characters we do not trust; N.B. changes in the characters should be based on biological evidence, do not forget this). After the characters have been changed we can re-analyse the matrix and evaluate the results again. (Most Americans like to order their characters a-priori, changes in the order of characters can also help to improve an analysis.)

ad. 2. Characters usually have an equal weight in an analysis, they are all regarded to be of equal value (in the computer programs the weight of every character is 1). We can give different weights to the characters, characters we trust (or have a high ci and ri) will get a high weight, others a low weight (e.g. a weight of 10 means that 2 changes in that character do not count as 2 steps, but as 2*10 = 20 steps). It will be very arbitrary if we provide the weights ourselves, it is better to use the rc for this purpose. Characters with a high rc receive a high weight, others a low weight. In the program Hennig86 we can use a reiterative weighting procedure (one procedure is the command **xsteps w;**, followed by **ccode;** to make the weights visible on the screen and by a new tree calculating command, e.g. **bb*;**). After each procedure we repeat it with the same commands till the weights of the characters do not change anymore. If lucky, one tree, of the many parsimonious ones, has been selected. Nasty side effects can be, that this tree was not among the original set of parsimonious trees, or that instead of several parsimonious trees many more will be found.

ad. 3. We can apply different criteria to select among the most parsimonious trees. Two options consider trees with the homoplasy concentrated in a few characters to be have

more deciseviness than equally parsimonious trees with the homoplasy spread over many characters. In these options a weight procedure is used to select the parsimonious trees with few homoplasious characters and if you are lucky this will just be a single cladogram:

-Goloboff (1993a) uses a weight factor F (for tree Fitness), which does not weight linearly like in Farris's program Hennig86, but concavely (much in accordance with the weighting guide lines issued by Farris in 1969). Unfortunately, this weight has the same disadvantage as the one in Hennig86, it might select trees not present among the most parsimonious trees (but see also Turner & Zandee, ms. or chapter 3.2.3.4 in Turner, 1995).

-Turner and Zandee (ms., see also chapter 3.2.3.3 in Turner, 1995) propose a different index based on the ri per character. This index still has to be evaluated.

-The most intriguing criterium is the application of the second law of thermodynamics to phylogeny (introduced by Brooks & Wiley, 1988). Instead of parsimony, entropy is used to select a tree. Geesink and Zandee developed the redundancy quotient (based on the information theory with algorithms parallel to those of thermodynamics), which is implemented in Zandee's computer program CAFCA (Zandee & Geesink, 1992; see below also). The most informative cladogram (highest rq) is selected, which is usually also among the most parsimonious trees, unless the latter contain many polytomies.

ad. 4. Different methods can be used to calculate a cladogram.

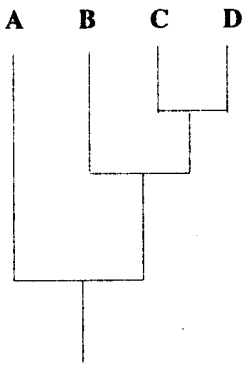
-Character compatibility. Only characters which are compatible with each other are used to construct a cladogram. However, when the cladogram contains many homoplasies, most characters are ignored or they are only used in secondary (or later) partial analyses. The result usually is a far from parsimonious cladogram.

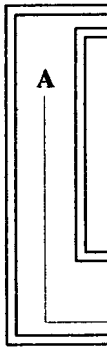
-Group compatibility. Applied by Zandee in his computer program Cafca (1994). Based on characters (or combinations of them) groups of taxa are identified. Those groups which are compatible with each other (completely including or excluding each other, not partly overlapping) are used to construct the cladogram. Again, if the level of homoplasy is high several groups cannot be recognized and a non-parsimonious cladogram will result (or when the most elaborate group searching criterium is used many trees will result).

-Maximum Likelihood Method. Advocated by Felsenstein (1982) and theoretically perhaps the best method. It is a statistical method, which uses the chances of character state changes to find a cladogram. Unfortunately, the chances of state changes are usually unknown and this method has only been tried in some molecular studies.

CLASSIFICATION

A cladogram or phylogenetic tree can be transformed in
shows a simple example. In principle every inclusive group, or put
be provided with its own rank. The largest group will receive the hi
of two or more taxa the lowest rank. The height of the rank is usu
has always been a family it will remain a family. Fig. 10 shows tha
groups formally, four ranks have to be used and, because the clad





Inclusive

Phylogenetic Classifications

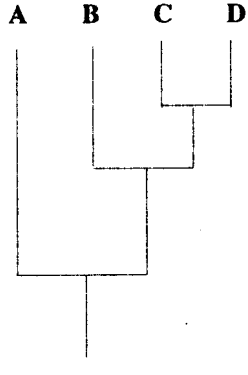
	Strict	Sequen
Family:	ADaceae	ADacea
Subfam.:	Aoideae	
Genus:	Aus	Aus
Species:	A	A
Subfam.:	BDoideae	
Genus:	Bus	Bus
Species:	B	B
Genus:	CDus	CDus
Species:	C	C
Species:	D	D

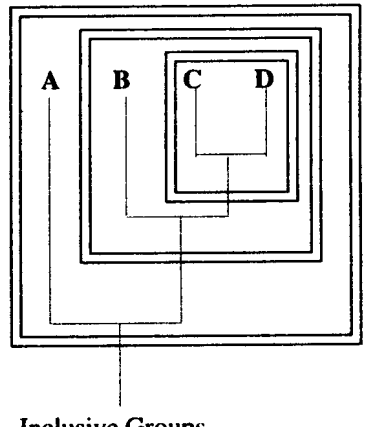
Fig. 10. Cladogram (top left) with its inclusive groups (top right), a

Phylogenetic Research: A Short Introduction

CLASSIFICATION

A cladogram or phylogenetic tree can be transformed into a classification. Fig. 10 shows a simple example. In principle every inclusive group, or put differently, every level will be provided with its own rank. The largest group will receive the highest rank, the smallest unit of two or more taxa the lowest rank. The height of the rank is usually a matter of history, if it has always been a family it will remain a family. Fig. 10 shows that if we treat all the inclusive groups formally, four ranks have to be used and, because the cladogram is asymmetric, many





Phylogenetic Classifications

	Strict	Sequence rule
Family:	ADaceae	ADaceae
Subfam.:	Aoideae	
Genus:	Aus	Aus
Species:	A	A
Subfam.:	BDoideae	
Genus:	Bus	Bus
Species:	B	B
Genus:	CDus	CDus
Species:	C	C
Species:	D	D

Fig. 10. Cladogram (top left) with its inclusive groups (top right), and two alternative classifications.

monotypic taxa are a result. Many problems are present to realize a classification from a cladogram. Wiley (1981) shows several rules which can be followed, one of them is the sequence rule of Nelson (1973), with which the number of ranks is reduced by giving the same rank to groups which branch off consequetively. The inclusiveness of the groups is then shown by the sequence in the classification, the first group includes the others. Usually, scientists hesitate to make a strict classification out of their cladogram, because it involves many extra ranks, often of groups which are difficult to recognize (and/or they do not trust their cladogram). However, the strictest adherence to the cladogram provides the classification with the highest information content.

NEW PERSPECTIVES

The phylogenetic analysis of DNA sequence data is becoming increasingly important and challenging. Several types of DNA are available, nuclear, mitochondrial, ribosomal, and (in plants) chloroplast DNA. Within the different types of DNA, parts are conservative in which a change/mutation may cause the death of the organism, while other parts are non-informative and may be very variable among the different specimens of a species. The former are often used in analyses on high classification levels (orders-phyla), while the latter are mainly used for analyses on (infra)specific level. The use of molecular data is important, but their value is perhaps somewhat overrated. Molecular analyses encounter more difficulties (and are therefore more challenging) than macromorphological analyses. The correct part of DNA has to be selected, conservative and variable enough for the problem to be solved. The different sequences have to be aligned, whereby similar parts are placed next to each other; when the DNA is very variable and shows many deletions this is very difficult and statistical methods have to be applied. Recognizing homology on the molecular level is much more complicated than on the macromorphological level (Patterson, 1988), just like selecting characters (is the complete DNA one character - Doyle 1992, or is a single base a character, or a triplet, or a change in one base, a single deletion, or a major deletion ?). Finally, phenetic and phylogenetic algorithms for calculating the cladogram are used, which make several analyses of more or less the same data incomparible. Phenetic analyses are used, because of the molecular clock idea. In every piece of DNA changes occur regularly and these are not mitigated by selection. Therefore, the amount of change in DNA is linear with time, which means that taxa very much alike only just recently developed, while others, with many differences, split up earlier. Consequently, similarity measures (phenetic measures) can be used to estimate the evolution of a taxon.

In conclusion, the analysis of molecular data takes a lot of time and energy, which increases the tendency to overestimate the value of the results. The latter should be evaluated against the - just as important - morphological evidence and the evidence of other parts of (other) DNA ('a gene tree need not be synonymous with a species tree' - Doyle and references therein 1992; see also Patterson *et al.*, 1993).

Phylogenetic Research: A Short Introduction

Cladistic methods are also applied to biogeographic data, whereby cladograms of areas, area(clado)grams are produced, showing the sequence in which the different areas split off after they have been united (e.g. during the Pangea period). Cladistic or vicariance biogeography does not stress the idea that taxa developed in a center of origin after which dispersal and subsequent speciation resulted in the taxa we know nowadays. Instead it uses the vicariance model as overall explanation for patterns. Vicariance patterns are probably the result of processes which influenced all taxa simultaneously (e.g., the separation of land after the Pangea period: ancestor in Pangea, its daughter species in the separate continents, Laurasia and Gondwana, etc.), while processes like dispersal and extinction are regarded as ad-hoc events, typical for a single taxon. The latter have to show up as homoplasies in the areagrams. Several taxa, occurring in the same areas, have to be analysed simultaneously, otherwise several ad-hoc occurrences will show as general vicariance events. Lately, several methods have been developed, component analysis (Page, 1994, after Nelson & Platnick, 1981), three area statements (Nelson & Platnick, 1991), component compatibility analysis (Zandee & Roos, 1987), Brooks' parsimony analysis (Wiley, 1988a, b), usually with same results as the former method), ancestral areas (Bremer, 1992). None of these methods is really superior and it is highly questionable whether areas may be used in the same way as taxa in a cladistic analysis.

Finally, ecological and behavioural data can also be used in cladistic analyses. Brooks and McLennan (1991) introduce this in a nice book which is relatively easy to read. Data can be incorporated in a (separate) data matrix or they can be evaluated on a cladogram which is based on other (morphological) data. The latter may be a help in evaluating different parsimonious cladograms. An example of the use of ecological data. Several species of a plant genus may occur at low altitude, while others are found in the mountains. One may wonder why one group occurs at low altitude and the other at high altitude. The cladogram may show that the occurence on low altitude is primitive and the presence in the mountains derived. This means that one of the questions was false, we do not have to wonder why the plants occur at low altitude, their ancestors already lived there. N.B. in another, as parsimonious cladogram, it may appear that the change to high altitudes occured twice, then this is an argument to reject this cladogram and to prefer the one with a single change.

FINAL REMARKS

Not all subjects concerning phylogenetic analysis have been evaluated nor elaborately treated. The basic concepts are explained. I like to give the following advice:

-if one is not an adept in phylogenetic analysis, then start to practise with small groups (5 taxa), the results are easier to understand and to interpret.

-always note in articles the characters and their states, the matrix and the results. Other people might not agree with you, but they can always repeat and understand your analyses and conclusions.

-be critical, do not believe the results of the computer at once. Evaluate the results yourself (perhaps with a program like Macclade) and use biological evidence to support conclusions or to change optimisations. Remember that when large groups are analysed the most parsimonious tree might not have been found.

-if the results are not up to expectation, e.g. many trees are found, then first start to re-evaluate your characters and their states again. Only use weighting and other methods once you are very certain of your characters.

-most phylogentic debates are found in the journals Cladistics and Systematic Biology (and to a lesser extent Systematic Botany). If one considers these to be too difficult, then first read a more elaborate introduction in Forey *et al.* (1992) and Wiley (1981).

Finally, do not think that phylogenetic analyses are too difficult and refrain from using them. Just do it and increase your experience by starting with small groups, discussing the results with fellow scientists in your institute, and by sending manuscripts to collegues for comments (they too consider it a challenge).

Acknowledgements

I am grateful to my wife, Naovarat, and the Late Prof. Sivarajan for comments and improvements of the manuscript. Hopefully the article has transformed into an enjoyable exercise.

Literature Cited

Bremer, K. 1992. Ancestral areas: a cladistic reinterpretation of the center of origin concept. *Syst. Biol.* **41**: 436-445.

Brooks, D.R. & D.A. McLennan. 1991. *Phylogeny, ecology, and behavior: a research program in comparative biology*. University of Chicago Press, Chicago.

Brooks, D.R. & E.O. Wiley. 1988. *Evolution as entropy. Toward a unified theory of biology*, 2nd edition. University of Chicago Press, Chicago.

Doyle, J.J. 1992. Gene trees and species trees: molecular systematics as one-character taxonomy. *Syst. Bot.* **17**: 144-163.

Farris, J.S. 1969. A successive approximations approach to character weighting. *Syst. Zool.* **18**: 374-385.

Farris, J.S. 1988. *Hennig 86 version 1.5 computer program and manual*. University of Stony Brook, New York.

Felsenstein, J. 1982. Numerical methods for inferring evolutionary trees. *Quarterly Rev. Biol.* **57**: 379-404.

Forey, P.L., C.J. Humphries, I.J. Kitching, ., R.W Scotland, D.J. Siebert, & D.M. Williams, 1992. *Cladistics, a practical course in systematics*. Clarendon Press, Oxford.

Goloboff, P.A. 1993a. Estimating character weights during tree search. *Cladistics* **9**: 83-91.

Phylogenetic Research: A Short Introduction

Goloboff, P.A. 1993b. *Pee-Wee, version 2.0. Computer program and manual.* Published by the author, New York.

Hennig, W. 1950. *Grundzüge einer Theorie der phylogenetischen systematik.* Deutsche Zentralverlag, Berlin.

Hennig, W. 1966. *Phylogenetic systematics.* University of Illinois Press, U.S.A.

Kornet, D.J. 1993a. Permanent splits as speciation events: a formal reconstruction of the internodal species concept. *J. Theor. Biol.* **164**: 407-435.

Kornet, D.J. 1993 b. *Reconstructing species. Demarcations in genealogical networks.* PhD thesis, Leiden University, Leiden.

Maddison, W.P. & D.R. Maddison. 1992. *MacClade, version 3.04.* Sinauer Associates, Sunderland.

Nelson, G.J. 1973. Classification as an expression of phylogenetic relationships. *Syst. Zool.* **22**: 344-359.

Nelson, G.J. & N.I. Platnick. 1981. *Systematics and biogeography.* Columbia University Press, New York.

Nelson, G.J. & N.I. Platnick. 1991. Three-taxon statements: a more precise use of parsimony? *Cladistics* **7**: 351-366.

Page, R.D.M. 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol.* **43**: 58-77.

Patterson, C. 1988. Homology in classical and molecular biology. *Mol. Biol. Evol.* **5**: 603-625.

Patterson, C., D.M. Williams & C.J. Humphries. 1993. Congruence between molecular and morphological phylogenies. *Annu. Rev. Ecol. Syst.* **24**: 153-188.

Pimentel, R.A. & R. Riggins. 1987. The nature of cladistic data. *Cladistics* **3**: 201-209.

Stevens, P.F. 1980. Evolutionary polarity of character states. *Annu. Rev. Ecol. Syst.* **11**: 333-358.

Stevens, P.F. 1991. Character states, continuous variation and phylogenetic analysis: a review. *Syst. Bot.* **16**: 553-583.

Swofford, D.L. 1993. *PAUP, version 3.1.1.* Smithsonian Institution, Washington D.C.

Turner, H. 1995. Cladistic and biogeographic analyses of *Arytera* Blume and *Mischarytera* Gen.nov. (Sapindaceae) with notes on methodology and a full taxonomic revision. *Blumea Add. Ser.* **6**: 1-230.

Turner, H. & M. Zandee. ms. The behaviour of Goloboff's tree fitness measure F. Accepted for *Cladistics.*

Wiley, E.O. 1981. *Phylogenetics: the theory and practice of phylogenetic systematics.* Wiley Interscience, New York.

Wiley, E.O. 1988a. Vicariance biogeography. *Annu. Rev. Ecol. Syst.* **19**: 513-542.

Wiley, E.O. 1988b. Parsimony analysis and vicariance biogeography. *Syst. Zool.* **37**: 271-290.

Zandee, M. 1994. *CAFCA - a Collection of APL Functions for Cladistic Analysis, PC computer program*

version 1.9.9a. Research Institute for Evolutionary and Ecological Sciences, University of Leiden, Leiden.

Zandee, M. & R. Geesink. 1992. *RQ. The redundancy quotient for cladograms, computer program version 1.0f.* Research Institute for Evolutionary and Ecological Sciences, University of Leiden, Leiden.

Zandee, M. & M.C. Roos. 1987. Component-compatibility in historical biogeography. *Cladistics* **3**: 305-332.

APPENDIX

The appendix contains a short manual for the program Hennig86 (Farris, 1988). It will not discuss all the commands possible in Hennig86, only the ones which will be most commonly used. The manual itself is very condensed and perhaps only readable for the well-versed cladist. A few items have to be pointed out:

- only use small case characters in the commands, no capitals.

- the program starts counting with 0, which means that if you want to change something in character 5 you have to refer to character 4 in your command. Likewise, tree 1 in the output will be the second tree.

- end every command with a semicolon (;). If you forget this the program just repeats the last command and stops any action. Typing the ; and pressing return will activate the program again.

- spaces and dots have special meanings, a space indicates that a new part of the command starts, while a dot means from...to (0.5 = character 1 to 6). Therefore do **not** use spaces and dots in the names of your taxa!

In this manual I suppose that your progam Hennig86 is in the subdirectory H86 of your computer. First we will use the MS-DOS text editor EDIT to make input (batch) files for the program Hennig86. Once those are finished we can start Hennig86, load the batch files and see which results we will get.

At start your computer will be in the base directory C:\. (In the rest of the program comment is typed between [ ], these are **not** part of the commands and one should **not** type them.)
Type: **cd\h86** and press return [change directory to C:\H86\]
Type: **edit exam1** and press return [make batch file called exam1]

Now we are in the MS-DOS text-editor, you can do the same with any wordprocessing program, but always save as ASCII- or DOS-file. Type the following lines, not the comments:

display*; [shows output on screen]
log c:\h86\exam1.out; [stores output in file EXAM1.OUT]
xread 'Example1 matrix with character 1 not split'
[this is a long command, between
' ' is some comment]

6 5	[number of characters and taxa]
outgroup	[name of first taxon]
111111	[characters of first taxon]
A	[name of second taxon, etc.]
212221	
B	
312232	
C	
121242	
D	
121252	
;	[end of the xread command]
ccode -.;	[make all characters unordered, - means unordered, . = all char; -3.6 8 = unorder 4 to 7 and 9]
ie*;	[calculate cladograms, use only when dataset is not complicated otherwise type: **mhennig*;bb*;**]

Now the file is ready, leave EDIT program (press ALT-key, F, X, Y). We can make the second file by typing **EDIT EXAM2**:

display*;

log c:\h86\exam2.out;

xread 'Example2 with character 1 split into 2 characters' 7 5 outgroup 1111111 A 2212221 B 2112232 C 1121242 D 1121252;

ccode -.;

ie*;

[Note that the xread command is condensed, also note that the number of characters is now 7.]

Next we can start Hennig86 by typing the command

SS and pressing return.

The program after having started shows a title and the cursor will be blinking (behind *> at the bottom line) and waiting for your first command. The first command will be loading the first batch file, once this is done, all commands in the file will be executed (output log file will be opened, matrix will be read, character will become unordered, and trees will be calculated).

Type: **proc exam1;** [do not forget the ;]

Peter C. van Welzen

Within a second the following output occurs on the screen:

xread

Example1 matrix with character 1 not split

ie length 11 ci 90 ri 66 trees 2

It repeats the xread comment and shows the result of the tree calculations: 2 trees are found, 11 steps long, the consistency index is 0.9 and the retention index is 0.66.

If we like to see the trees we have to give an extra command, also if we like to see the statistics belonging to the tree. These two commands could have been incorporated in your batch file already, but if many trees are found, this will create an enormous output file. Usually, if many trees are found, it will be a complete waste to see the trees, then it is better to leave the program and change the characters.

To see trees type (within the Hennig86 program): **tplot;**

and for the statistics type: **xsteps hclm;**

If you like to see the output, leave the program by typing

yama (according to some the name of an Indian God?).

The output can be studied through a wordprocessor or text editor. Look for the file **EXAM1.OUT** (not EXAM1, this was the input file!). The complete output looks like:

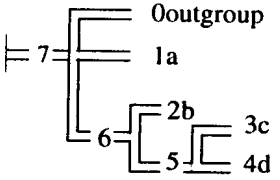
xread

Example1 matrix with character 1 not split

ie length 11 ci 90 ri 66 trees 2

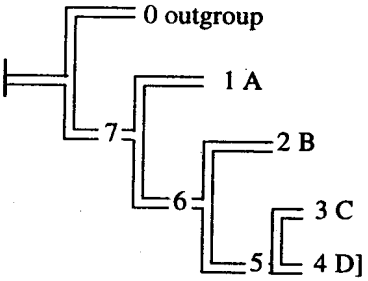
tplot file 0 from ie 2 trees

tree 0

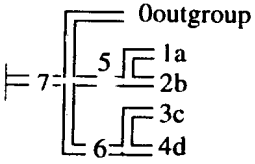


[Result of tplot: 2 trees depicted]

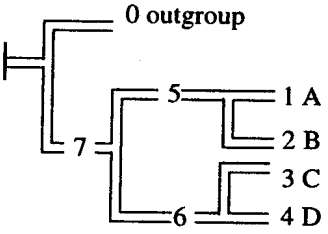
[Change into



tree 1



[Change into



Phylogenetic Research: A Short Introduction

```
      xsteps file 0 from ie 2 trees                [Result of xsteps command]
      tree 0                                       [Results for first tree]
character 0                                        [Character states of character 1
      5    6    7                                  on nodes 5, 6 and 7 of upper
      1    1    1                                  tree: no change]
character 1                                        [Character states of character 2
      5    6    7                                  one change on node 5 from state
      2    1    1                                  1 to state 2]
character 2                                        [Character states of character 3
      5    6    7                                  one change, unknown on node 6 or
      1   12   12                                  7, change from state 2 to 1]
character 3
      5    6    7                                  [No change on internal nodes]
      2    2    2
character 4                                        [All taxa have another state,
      5    6    7                                  undecidable which change where,
     12+ 12+  12+                                  12+ means all possible: 1-5]
character 5
      5    6    7                                  [Change on 6 from state 2 to 1]
      2    2    1
tree 0 length 11 ci 90 ri 66                       [Summary of this tree:]
character/steps/ci/ri
       0   1   2   3   4   5                       [Character numbers: 1-6]
       2   1   2   1   4   1                       [Changes per character]
     100 100  50 100 100 100                       [Consistency index per character]
     100 100   0 100 100 100                       [Retention index per character,
                                                   char. 3=0, no synapomorphies]

      tree 1                                       [Same statistics for second tree]
character 0
      5    6    7
    123    1    1
character 1
      5    6    7
      1    2    1
character 2
      5    6    7
      2    1    1
character 3
      5    6    7
      2    2    2
```

```
character 4
        5     6     7
       12+   12+   12+

character 5
        5     6     7
       12     2    12

tree 1 length 11 ci 90 ri 66
character/steps/ci/ri
        0  1  2  3  4  5
        2  1  1  1  4  2
      100 100 100 100 100  50
      100 100 100 100 100   0
```

best fits [Summary of best and worst fits]

```
character/steps/ci/ri
        0   1   2   3   4   5
        2   1   1   1   4   1
      100 100 100 100 100 100
      100 100 100 100 100 100
```

worst fits

```
character/steps/ci/ri
          0   1   2   3   4   5
          2   1   2   1   4   2
        100 100  50 100 100  50
        100 100   0 100 100   0
tree/length
          0   1
         11  11
```

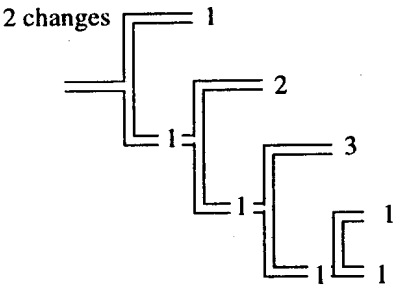
It will be up to us to interpret these data correctly. First of all we have to change the trees somewhat. Hennig86 always provides a trichotomy at the base of the tree, which has to be changed in two dichotomies with the outgroup at the base of the cladogram, outside of our ingroup. Then we have to find where in the trees the characters change, which is a somewhat complicated procedure. We can, within the program Hennig86, use the XX command, but I will not explain this user unfriendly command here. We will use a combination of the XSTEPS statistics, our matrix, and the tree to find the changes in the tree. Of the XSTEPS statistics we need the overview at the end of each tree (with the character numbers, changes, ci, and ri per character) and the states on the internal nodes 5-7 per character. In the tree 0-4 are the terminal nodes, i.e. the taxa in our data matrix, 5-7 are nodes the computer found in the most parsimonious cladogram and these are called the internal nodes. We duplicate our tree as many times as we have characters (per character we will have to write all the states on every node before we can decide where the change occurs, therefore one tree per character is needed), and we add one more tree, the one in which we summarize our

Phylogenetic Research: A Short Introduction

results, this will be the tree we will publish. The character states for the different nodes we find in our data matrix for our taxa (terminal nodes 0-4) and for the internal nodes in the XSTEPS output.

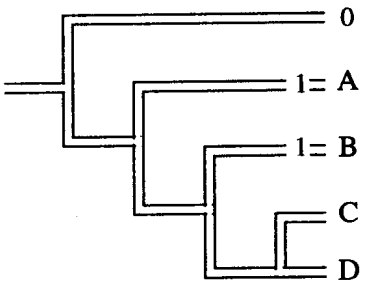
Tree 1 (Tree 0 of Hennig 86), character 1

Tree overview States Change Endresult



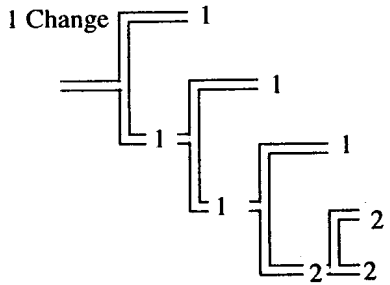
A: 1 to 2

B: 1 to 3

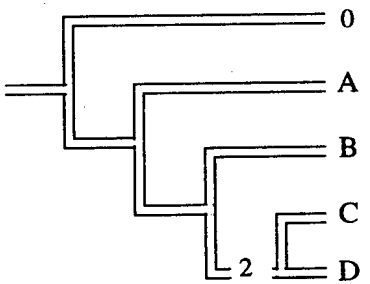


All internal nodes have the same state, therefore the changes must occur in terminal nodes. We have two autapomorphies.

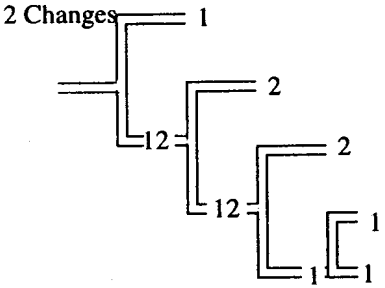
Character 2



node 5 : 1 to 2



Character 3

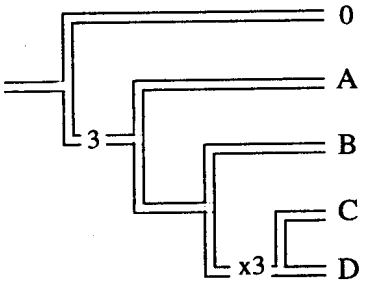


A : 1 to 2

B : 1 to 2 or

node 7 : 1 to 2

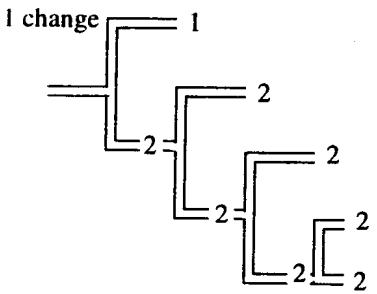
node 5 : 2 to 1



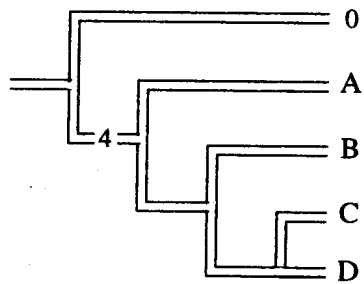
Peter C. van Welzen

Here we have the ACCTRAN/DELTRAN choice, either we interpret this character as a parallel development (autapomorphies in A and B) or as a reversal (synapomorphy at 7 and another one, the reversed, at 5). I opt for the reversal (entire-crenate-entire leaflets, instead of two times the origin of crenate).

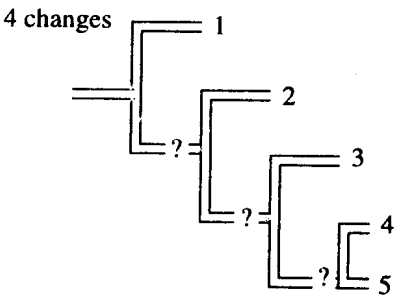
Character 4



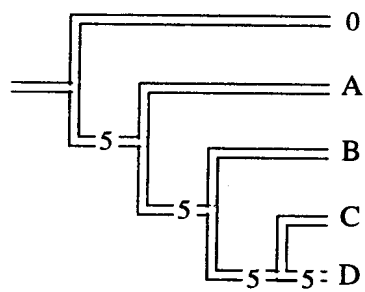
node 7:1 to 2



Character 5



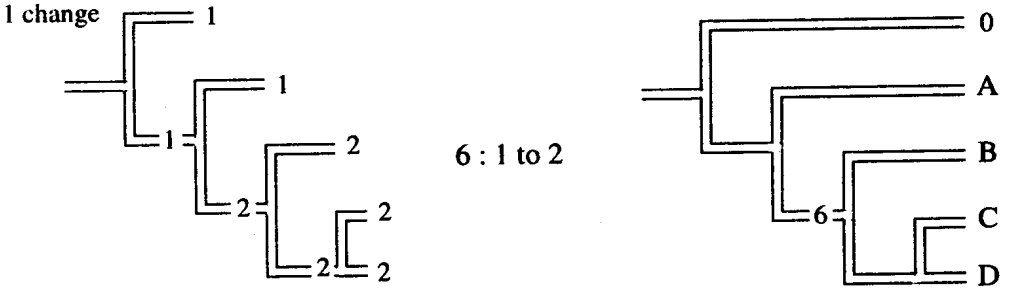
7 : 1 to 2
6 : 2 to 3
5 : 3 to 4
D : 4 to 5



Here all changes are possible, this character can be interpreted as all autapomorphies, or, as shown above, as deep synapomorphies or every solution in between. All optimisations are as parsimonious, and biological evidence does not provide a choice also. This character should be omitted from the analysis.

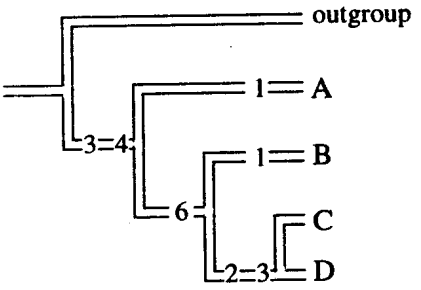
Phylogenetic Research: A Short Introduction

Character 6

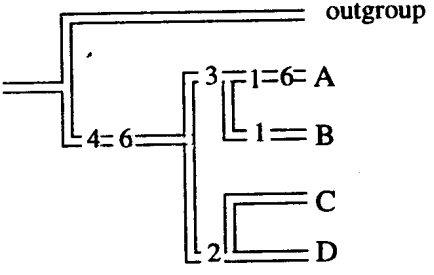


Final result:

Tree 1 (0 of Hennig 86)



Tree 2 (1 of Hennig 86)



Try the second tree yourself; here character 6 is creating the homoplasy. We have to select one of the two trees. One of the options is to change our matrix (as done in EXAM2) or we can start character weighting. For the latter I will briefly give the commands, which should be given when the program Hennig86 is active (before you type YAMA):

xsteps w;	[calculates weights per character]
ccode;	[not necessary, but shows weights]
bb*;	[or **ie***;, calculates trees with new weights]

Continue this procedure till the weights do not change anymore. In this example weighting does not help.

RESULTS OF EXAM2:

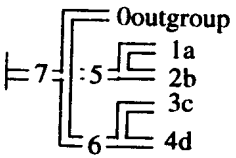
xread

Example2 with character 1 split into 2 characters

ie length 11 ci 90 ri 75 trees 1

tplot file 0 from ie 1 tree

Peter C. van Welzen



xsteps file 0 from ie 1 tree

tree 0

character 0

5	6	7
2	1	1

character 1

5	6	7

character 2

5	6	7
1	2	1

character 3

5	6	7
2	1	1

character 4

5	6	7
2	2	2

character 5

5	6	7
12+	12+	12+

character 6

5	6	7
12	2	12

tree 0 length 11 ci 90 ri 75

character/steps/ci/ri

0	1	2	3	4	5	6
1	1	1	1	1	4	2
100	100	100	100	100	100	50
100	100	100	100	100	100	0

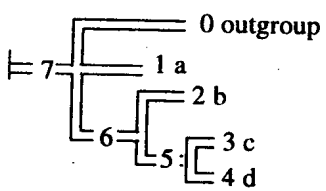
tree/length

0
11

ie length 12 ci 83 ri 66 trees 2

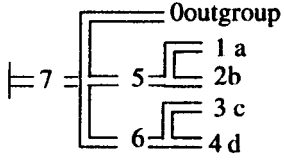
tplot file 0 from ie 2 trees

tree 0



Phylogenetic Research: A Short Introduction

tree 1



The first part must be clear. Only one tree is found (the more symmetric tree of our first analysis) and the statistics of it are provided. It must be possible to optimize the characters yourself. At the end I have given, after the xsteps results, an extra command: **ccode +5;**, followed by the **ie*;** command again. The ccode changed character 6 (petal colours) in an ordered character (indicated by the plus). The ie command calculated with this and the same two trees of our first analysis are found again. See text for an explanation.